Response to comments on "Two-tailed significance tests for 2 × 2 contingency tables

Response to comments on "Two-tailed Significance Tests for 2x2 Contingency Tables: What is the Alternative?"

Robin John Prescott

Centre for Population Health Sciences, Usher Institute, University of Edinburgh, UK
robin.prescott@ed.ac.uk

Keywords

2x2 Contingency Tables

Two-tailed tests

Alternative hypotheses

Fisher's exact test

“What is the alternative?” That was the question asked as part of the title of my paper[1]. This was not mere whimsey on my part, nor was it a rhetorical question. It was a subtle play on words as a reminder of my key point that the inferential methods used to analyse a 2x2 contingency table should depend on the Alternative Hypothesis to the Null that is chosen. Apparently, I was too subtle for my own good. The correspondence from Martín Andrés and colleagues[2] indicates that they missed this point. Once this is appreciated, the consequences should be clear. The choice of one Alternative Hypothesis (independence of the rows and columns of the 2x2 table) leads to one set of analytical methods, and these are the ones most commonly used in medical research practice. It also leads to the methods advocated by Martín Andrés and colleagues. On the other hand, adopting the Alternative Hypothesis that either group A is superior to group B or vice versa, using the ICH definition of superiority, (or simply applying a policy of symmetry, which is the standpoint of many), leads to a different set of methods.

I emphasised that under the first of the aforementioned Alternative Hypotheses, the use of methods such as Fisher’s added-tails test would be preferable because of their added power and I mentioned that I would not hesitate to use them in these circumstances. However, I developed the argument that in medical research “if we are applying a two-tailed test of significance, we should be seeking to differentiate between three possible conclusions with our inferences, rather than a simple binary decision of acceptance or rejection of the Null hypothesis. If we reject the Null hypothesis we should either be able to claim that one group is superior to the other [in the sense of the ICH guidelines] or vice versa”. More generally, I suggested that in medical research, “for any asymmetrical test, a composite alternative hypothesis based on superiority of one or other group should be used, effectively leading to the application of two one-tailed tests at the $\alpha/2$ level”.

In the context of the correspondence from Martín Andrés and colleagues, the final paragraph of the paper bears repeating: “It must be recognised that the methods advocated in this paper come with a downside. The downside is that power is reduced compared to using a general alternative to the Null, and arguments about power and the size of the test have been central to many of the arguments about the best way to analyse 2x2 contingency tables. Despite this, the author believes that this is a necessary price, well worth paying, in order to obtain coherent inferences in medical research”.

It seems therefore that Martín Andrés and colleagues have not picked up on these points. Nowhere in their letter do they mention choice of an Alternative Hypothesis. They conclude, wrongly, that my objective was to avoid paradoxical conclusions. That is certainly a consequence of my recommendation for medical research papers but my hope was that recognition that different choices of Alternative Hypotheses, that lead to different choices of methods, would help reconcile some of the disharmonious thoughts on the analysis of 2x2 tables. It is not wrong to analyse a r x c contingency table using a chi-squared test with (r - 1) x (c - 1) degrees of freedom, nor is it wrong to analyse using the Jonckheere-Terpstra test. Similarly, I would argue that for 2 x 2 tables, both the Fisher’s added-tails and the Fisher’s Double P-value, as examples, are appropriate under different Alternative Hypotheses. The correspondents seem to suggest that use of the Fisher’s Double P-value is always contraindicated – a position with which I fundamentally disagree. My implicit plea for acceptance of different points of view according to the choice of the Alternative Hypothesis may have fallen on deaf ears.

I think it is worth pointing out that the views expressed by Martín Andrés and colleagues are not necessarily always correct.

The initial criticism of Martín Andrés and colleagues is concerned with power and how researchers will have to design their experiments with sample sizes that are too large. This criticism is misplaced. As I have already emphasised, the point about power was made very clearly in the original publication, both in the Introduction as well as the final paragraph which is reproduced above. In my view, the sample size should depend on the planned Alternative

Hypothesis. If researchers are working in an area outside clinical medical research, the general alternative to the Null may be a perfectly reasonable choice and standard sample size calculations will apply. If you believe in showing superiority as the ICH defines it, or are simply an advocate of symmetrical testing, then the sample sizes should be marginally larger.

It is interesting that Martín Andrés and colleagues comment that my recommended test (Fisher's Double-P value) is the least powerful test possible. That statement depends on the Alternative Hypothesis. The conservatism of the one-sided Fisher's exact test is due solely to its granularity. That is, its p-value is generally unable to exactly equal a specified value of α/2. If that granularity is removed by Tocher's approach[3] (which, I should add, would not be used in practice), it is the uniformly most powerful unbiased test.

In paragraph 3, Martín Andrés and colleagues suggest that doubling the p-value of the one-tailed test has been defended by "other (few) authors". This is not so. As I mentioned in the paper, this was the approach favoured by Fisher himself. Other more contemporary authors, such as Bland[4] and Campbell and Machin[5] come to the same conclusion. In The Analysis of Binary Data, Cox[6] expresses a similar view in more mathematical notation. His equation (4.17) expresses the two-tailed p-value as double the minimum of the two one-tailed probabilities and states that "it is usual to quote the value" in this way.

Much of the rest of the rest of the correspondence is directed to showing that by taking the general alternative to the Null and by using asymmetrical confidence intervals, the examples I showed can be presented without any paradoxical results. Under that Alternative Hypothesis, I agree with much, but not all, of what has been written but, as stated earlier, it is missing the point. The paradoxes that I identified in my paper occur only when considering the composite alternative hypothesis that one of the two treatments is superior to the other.

Martín Andrés and colleagues then introduce Mantel's chi-squared method as a better approximation to the Fisher's added-tails test. I agree that this is the case, but it was not one of the four tests I was considering. The Pearson Chi-squared test only provides a good approximation at larger sample sizes as indicated in Tables 3 and 4 of the original paper. In fact, Mantel changed his view on the analysis of 2x2 tables in 1990[7]. He stated "Here I must recant a position I have previously taken in this situation. In fact I now have in preparation a report justifying the position taken by Yates that the two-tailed probability should routinely be obtained from a doubling of the single tailed probability".

Discussion in the letter then proceeds to unconditional tests. I had sought to avoid introducing this kind of complexity into the argument by limiting coverage to four common methods that illustrate the consequences of different alternative hypotheses. However, as the topic has been introduced it is worth making the point that there are eminent statisticians who disagree vehemently with the views expressed by Martín Andrés and colleagues. Mantel[7] states "In the event, and in my opinion, the current faulting[8] and earlier faulting or dispraise by others, of the exact test and its continuity-corrected approximation, reflects a lack of understanding of statistics or an appreciation of the conditioning principle, by those who find fault".

Martín Andrés and colleagues make the claim that after using conditioned methods, the parameter of interest can only be the odds ratio. I think this a failure to appreciate the conditioning principle alluded to by Mantel.

The correspondents also refer to the Barnard order[9] method. This method was later refuted by Barnard himself in favour of Fisher's exact conditional test[10]. Barnard stated "further meditation has led me to think that Professor Fisher was right after all".

Of the four methods that I examined, the most commonly used that I encounter while reviewing for medical journals is undoubtedly the Pearson chi-squared test. I found it interesting that the correspondents reject the use of this test, even under the general alternative hypothesis. At least that is something we can agree upon. Nevertheless, the correspondents will be well aware that there is a massive literature supporting the use of this test. This was made clear to me when an author produced such a list to repudiate my criticism of the use of the chi-squared test in a paper I was reviewing. I reference Haviland[8] here as one example, as it is in response to this paper that Mantel made his earlier referenced comments. I believe that until there is clear guidance to do otherwise, it will continue to be used on a regular basis in medical research papers because of its greater power.

Therefore, I state again my view that the medical statistics community needs to find something approaching a common position in producing recommendations to medical journals on suitable ways of summarising 2x2 tables. Even if it was unclear to some initially, on second reading of my paper, I hope it will be accepted that choice of the Alternative Hypothesis is critical in deciding which methods are to be recommended. For the specific case of medical research, I propose that we should be using methods that are compatible with the ICH guidelines. That leads to a two-tailed procedure that is equivalent to two one-sided tests at the $\alpha/2$ level and to the employment of the Fisher's Double-P value method and associated confidence intervals. Perhaps that can, at least, be a starting point to developing guidelines, or a contribution to the discussion if guidelines are already being proposed. I invite everyone with views on how this can be taken forward to contact me.

References

1. Prescott RJ. Two-tailed significance tests for 2 × 2 contingency tables: What is the alternative? Statist Med. 2019; 38:4264-4269.

2. Martín Andrés A, Herranz Tejedor I, Gayá Moreno F. Comments on “Two-tailed significance tests for 2 × 2 contingency tables: What is the alternative?”. Statist Med 2109; In press.

3. Tocher KD. Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika 1950; 37:130-144.

4. Bland M. An Introduction to Medical Statistics. 4th ed. Oxford, Oxford University Press; 2015.

5. Campbell MJ, Machin D. Medical Statistics A Commonsense Approach. 3rd ed. Chichester, John Wiley & Sons; 1999.

6. Cox DR. Analysis of Binary Data. London, Methuen; 1970.

7. Mantel N. Comment. Statist Med.1990; 9:369-370.

8. Haviland MG. Yates's correction for continuity and the analysis of 2 × 2 contingency tables. Statist Med. 1990; 9:363-367.

9. Barnard GA. Significance tests for 2x2 tables. Biometrika 1947;34:123-138.

10. Barnard GA. Statistical inference. J Royal Stat Soc Ser B 1949;11:115-139.